

Invited Commentary

Invited Commentary: Genes, Environment, and Hybrid Vigor

Marta Gwinn, Idris Guessous, and Muin J. Khoury

Initially submitted March 6, 2009; accepted for publication May 11, 2009.

In the 1950s, case-control studies of smoking and lung cancer established a paradigm for epidemiologic studies of risk factors for chronic diseases. Since then, thousands of case-control studies have examined possible associations of countless risk factors with numerous diseases, rarely finding associations as strong or consistent as that of smoking with lung cancer. Recently, researchers have applied advances in molecular genetics to conduct candidate gene and genome-wide association studies of lung cancer. Skeptics among both epidemiologists and geneticists have argued that genomic research adds little value when most cases of disease can be attributed to a preventable exposure; however, well-conducted studies of gene-environment interactions that draw on data from more than 50 years of research in toxicology, pathophysiology, and behavioral science offer important models for the development of more comprehensive approaches to understanding the etiology of chronic diseases.

case-control studies; DNA damage; DNA repair; genetic predisposition to disease; lung neoplasms; oxoguanine glycosylase 1, human; smoking

Abbreviations: *hOGG1*, human 8-oxoguanine DNA *N*-glycosylase 1; HuGE, human genome epidemiology.

The 20th-century spread of cigarette smoking led to a global epidemic of tobacco-related diseases, which in 2000 killed more than 4 million people worldwide (1). The yearly toll is expected to surpass 6 million by 2015; by then, approximately one-fifth of all tobacco-related deaths will be caused by lung cancer, which has become the leading cause of cancer mortality (1–3). Trends in lung cancer incidence and mortality tend to parallel trends in smoking prevalence, lagged by 20–30 years. In the United States, lung cancer incidence among men has declined since reaching a peak in the 1980s, and incidence among women has begun to plateau (3). However, the prevalence of smoking continues to grow in many countries, including China, which is the world's leading tobacco producer and is home to approximately one-third of the world's smokers, mostly men (4, 5). The public health burden of smoking-related diseases is shifting to the developing world, where lung cancer incidence can be expected to increase for decades to come (2, 6).

Although most lung cancers occur in smokers, most smokers do not develop lung cancer (7, 8). Potential mod-

ifiers of risk include environmental and occupational exposures, as well as diet, exercise, and other lifestyle factors (5). Genetic differences in susceptibility have long been suspected, because lung cancer risk aggregates in families, even when smoking behaviors are accounted for (9). Lung cancer has also been associated with several inherited cancer syndromes caused by rare, germ-line mutations (10). For more than 25 years, epidemiologists have studied the role of genetic variation in enzymes involved in the metabolism of tobacco carcinogens in relation to lung cancer (11). Recently, researchers have applied advances in molecular genetics to investigate an expanding set of candidate genes, particularly those involved in carcinogen activation and detoxification and repair of DNA damage (12). Data obtained on May 1, 2009, from HuGE [Human Genome Epidemiology] Navigator (www.hugenavigator.net), a continuously updated knowledge base in human genome epidemiology (13), indicate that since 2001, nearly 700 articles (including 29 meta-analyses) have examined 530 genes for a possible association with lung cancer. A systematic review of candidate cancer susceptibility genes found 11 variants of 10

genes associated with lung cancer (14); none of these appeared in 7 published genome-wide association studies (15–21), which identified additional loci for further investigation (www.hugenavigator.net).

The failure of genetic association studies and meta-analyses of published data to identify coherent patterns is not unique to lung cancer. Commenting on the largely null results obtained from the Breast Cancer Consortium's large-scale reanalysis of candidate single nucleotide polymorphisms, John Ioannidis suggested that success is unlikely without prospective collaboration among investigators and that "making progress on the genetics side [will] require progress on environmental components" as well (22, p. 1352). Against this background, the study of lung cancer, smoking, and human 8-oxoguanine DNA *N*-glycosylase 1 (*hOGG1*) genotype reported by Chang et al. (23) in this issue of the *Journal* is worth a closer look.

At least 20 previous studies have examined lung cancer in association with variants of the *hOGG1* gene, which encodes the *hOGG1* DNA repair enzyme, including 4 meta-analyses of the *hOGG1 Ser326Cys* variant. The most recent meta-analysis, a pooled analysis of 4 studies conducted by members of the International Lung Cancer Consortium (24), found a modest association in Caucasians (for *Cys/Cys*, odds ratio = 1.34, 95% confidence interval: 1.01, 1.79; reference group not specified) but not in Asians; analyses stratified by smoking status were not presented. A second meta-analysis, published at almost the same time (25), included 17 studies and found a significant association with lung cancer only in Asians (for *Cys/Cys* + *Ser/Cys* vs. *Ser/Ser*, odds ratio = 1.18, 95% confidence interval: 1.01, 1.38); only 4 studies provided sufficient information on smoking to be included in a stratified meta-analysis, where the association persisted only in nonsmokers.

In the current study, Chang et al. (23) recruited cases and controls from 6 medical centers in Taiwan as part of an ongoing, multicenter study of the genetic epidemiology of lung cancer. This study had several advantages for examining *hOGG1*-smoking interactions: a large size (more than 1,000 persons in both case and control groups), a lack of ethnic heterogeneity, high prevalences of the variant *Ser/Cys* and *Cys/Cys* genotypes (approximately 85% combined), and a high prevalence of smoking. By presenting genetic associations stratified by lung cancer cell type and smoking status, Chang et al. allow readers to assess for themselves the independent and joint effects of genotype and exposure on disease risk, rather than asking them to rely on modeled odds ratios or *P* values alone (26). When results for all lung cancers were stratified by smoking status, the only statistically significant association was with the *hOGG1 Cys/Cys* genotype among heavy smokers. Considering the joint effects of smoking and *hOGG1* genotype on lung cancer in a multiway table—stratified by smoking status (never, moderate, heavy), *hOGG1* genotype (*Ser/Ser*, *Ser/Cys*, *Cys/Cys*), and lung cancer cell type (adenocarcinoma, squamous cell, small cell)—the authors discovered a more general pattern of dose-dependent synergy (see Chang et al.'s Table 4 (23)). This pattern can be appreciated in a visual display of results for the adenocarcinoma subgroup, which accounted for 60% of cases (Figure 1). As the authors noted, failure to account

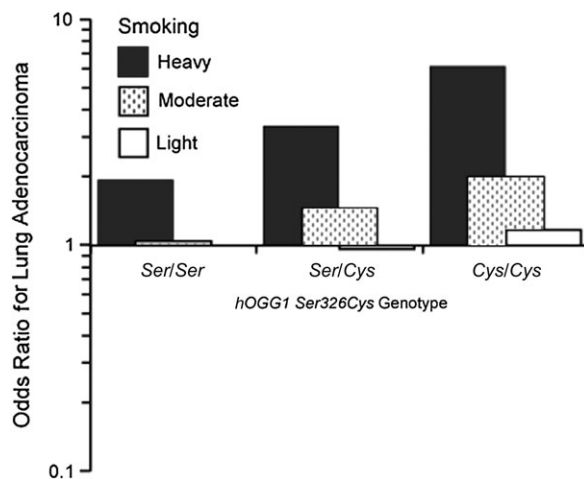


Figure 1. Interaction of the human 8-oxoguanine DNA *N*-glycosylase 1 (*hOGG1*) *Ser326Cys* polymorphism with smoking in adenocarcinoma of the lung, Genetic Epidemiological Study of Lung Adenocarcinoma, Taiwan, Republic of China, 2002–2004. Data were obtained from the paper by Chang et al. (23). For moderate smokers with the *Cys/Cys* genotype, $P < 0.05$; for heavy smokers with the *Ser/Cys* and *Cys/Cys* genotypes, $P < 0.001$.

for smoking in the analysis of this association would have masked the results.

It seems obvious that collecting and analyzing data on "candidate exposures" should be considered an essential component of "candidate gene" studies, especially when these exposures have been as thoroughly investigated as smoking. In contrast, most genome-wide association studies are as agnostic of exposures as they are of genes: Of nearly 300 genome-wide association studies indexed in the HuGE Navigator knowledge base, only 14 (5%) described potential gene-environment interactions (unpublished data from HuGE Navigator). Of these, 11 were pharmacogenomic studies (with a drug as the environmental factor), and just 1 was a study of lung cancer and smoking (15). Of the 6 other genome-wide association studies of lung cancer, 5 treated smoking history as a potential confounder (by either matching or adjusting for smoking) and 1 did not consider smoking at all. Commenting on the results of such studies, Wacholder et al. (27) observed that without smoking information, one cannot distinguish among variants that might influence lung cancer risk through effects on smoking behavior, effects on carcinogenesis, or both.

Why study genetic associations with lung cancer at all, when most cases can be attributed to a preventable exposure? Skeptics have argued that genomic research adds little value to population-level interventions for diseases with known environmental causes (28, 29). Carlsten and Burke challenged such research directly, asserting that "given the obvious dangers of tobacco and the associated imperative to eliminate it, research undertaken purely to unravel mechanisms of tobacco-related cancer is difficult to justify" (29, p. 2481). A general argument for gene-environment research on common diseases is that it provides insights into disease processes at the population, individual, and molecular

levels (30). Reducing uncertainty about such questions as etiologic fraction, natural history, and molecular mechanisms of disease is fundamental to developing efficient strategies for prevention, treatment, and further research. Although many of these questions have been examined extensively for lung cancer, the resulting approaches to prevention, screening, and treatment have had limited success. The most important approach—preventing smoking initiation—faces challenging social and economic obstacles; the second most important—smoking cessation—also faces a biologic obstacle (addiction). Furthermore, neither of these approaches reduces the risk of lung cancer in ex-smokers or in those who continue to smoke. Thus, additional goals of genomic research on smoking-related lung cancer are to assess individual susceptibility, develop more effective methods for smoking cessation, and prevent the evolution of cancer in former smokers (31). Using combinations of risk factors, including genetic susceptibility, to identify very high-risk groups for chemoprevention and screening trials is a priority (32).

An important rationale for well-conducted research on smoking and lung cancer is that it provides an opportunity to study the archetype of gene-environment interaction in epidemiology in the context of data from more than 50 years of research in toxicology, pathophysiology, and behavioral science. Attempts to understand and integrate such data often reveal the disconnections and deficiencies in current research approaches. Indeed, measuring genotypes may be the easiest part of gene-environment interaction research; environmental data are much more difficult to collect and analyze, for reasons recently summarized by Khoury and Wacholder (33). The collection of specific information on tobacco consumption, including duration and intensity, is often suboptimal in epidemiologic studies (34). At the molecular level, tobacco smoke presents a composite exposure of more than 3,500 particulate exposures and 500 vapor exposures, each with its own effects (35). Analysis of even relatively simple gene-environment interactions may not be straightforward; for example, interactions of metabolic genes with smoking may be nonlinear, varying with the amount of smoking exposure (34).

A careful review of previous studies is important for developing the next wave of epidemiologic research on lung cancer, which will look beyond single gene associations to complex interactions between multiple genes and environmental exposures. Chang et al. present genotype-specific risks, stratified by smoking status and lung cancer cell type (23); however, even this most fundamental presentation of results is missing from many published studies of associations of DNA-repair gene polymorphisms with lung cancer (12). Many small studies examine multiple potential interactions without having sufficient statistical power to detect them, leading to false-positive as well as false-negative results; however, lack of common measurement approaches makes it difficult to combine results from such studies for meta-analysis (36). These problems can only be addressed in part by the formation of consortia for conducting pooled analyses of individual-level data, such as the International Lung Cancer Consortium and the Genetic Susceptibility to Environmental Carcinogens Study (24, 34). New population-

based case-control studies, such as the Environment and Genetics in Lung Cancer Etiology Study, have been developed with the aim of integrating multiple epidemiologic, molecular, and clinical factors into a more complete model of lung cancer etiology (37).

Smoking has been studied for possible associations with a wide array of adverse health effects and remains a prominent topic in epidemiologic research. Since 1965, the *American Journal of Epidemiology* has published more than 1,700 articles that included the word “smoking” in either the title or the abstract; of these, 570 (nearly one-third) have appeared since 2000. Only 25 (4%) of these articles analyzed genetic variants in relation to smoking or smoking-related health outcomes, of which almost half ($n = 11$) were HuGE reviews (38). Carlsten and Burke’s criticism of genetic studies of lung cancer could just as well have been directed at epidemiologic studies of smoking, “given the obvious dangers of tobacco”; nevertheless, such research is important in defining the role of elusive environmental exposures (e.g., passive smoking) in conditions with incompletely understood causes (e.g., low birth weight)—a setting where measuring genetic polymorphisms may also be helpful (39). Sometime soon, we might expect studies of candidate exposures (“risk factors”) to consider candidate genes as well.

Case-control studies of smoking and lung cancer put “risk factor epidemiology” on the map at around the same time that Watson and Crick were describing the structure of DNA (8). Since then, thousands of case-control studies have examined countless risk factors for association with numerous diseases; however, few epidemiologic associations have been as strong and consistent as that of smoking and lung cancer. In the mid-1990s, epidemiology experienced an existential crisis regarding the value of “black box” methods, which seemed increasingly focused on the analysis of risk factors with ever-smaller effects (40). During the same period, advances in human genomic science and technology inspired a new wave of case-control studies conducted by geneticists seeking to identify genetic risk factors for common diseases. Moving quickly from candidate genes to genome-wide associations, this research has effectively reinvented “black box epidemiology” in a parallel universe where all of the risk factors are genetic polymorphisms. Of more than 40,000 genetic association studies published since 2001, fewer than 6,000 (16%) considered any potential interactions with environmental factors (unpublished data from HuGE Navigator).

After 50 years of convergent evolution in epidemiology and genetics, we should be seeking “hybrid vigor” in a more integrated approach to population-level research on human health and disease. Rather than focusing ever more narrowly on individual risk factors, epidemiologists should reclaim their pivotal role in integrating data from all levels—molecular, behavioral, environmental, social, and even planetary—to understand human health (41). Inaugurating an era of gene-environment-wide interaction studies, or “GEWIS,” will surely require unprecedented investment in large, collaborative studies that are able to collect and maintain the integrity of complex data in a way that permits multidimensional analysis (33). As new methods arise for identifying, measuring,

analyzing, and integrating such data, epidemiologists should be ready to use them. The analysis of gene-environment interactions is fundamental, but new research can expand the horizon for action even on problems as old and intractable as nicotine addiction (42). Finding order and direction in this universe will be challenging (43), but in the words of Neil Pearce, “If complexity is the price of being relevant and addressing the major public health problems, then so be it” (44, p. 715).

ACKNOWLEDGMENTS

Author affiliations: Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia (Marta Gwinn, Idris Guessous, Muin J. Khoury); and Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia (Idris Guessous).

This work was supported in part by the appointment of Dr. Idris Guessous to the Research Participation Program at the Centers for Disease Control and Prevention. The Research Participation Program is administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the US Department of Energy and the Centers for Disease Control and Prevention.

The authors thank Dr. Wei Yu and Melinda Clyne for development and curation of the HuGE Navigator knowledge base and Anja Wulf for help with graphics.

The findings and conclusions presented in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Conflict of interest: none declared.

REFERENCES

- World Health Organization. *The Tobacco Atlas*. Geneva, Switzerland: World Health Organization; 2002. (http://www.who.int/tobacco/statistics/tobacco_atlas/en/). (Accessed May 1, 2009).
- Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030 [electronic article]. *PLoS Med*. 2006;3(11):e442.
- Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008. *CA Cancer J Clin*. 2008;58(2):71–96.
- Food and Agriculture Organization of the United Nations. *Issues in the Global Tobacco Economy. 3 Tobacco in the People's Republic of China*. Rome, Italy: Food and Agriculture Organization of the United Nations; 2003. (<http://www.fao.org/docrep/006/y4997e/y4997e0g.htm>). (Accessed May 1, 2009).
- Molina JR, Yang P, Cassivi SD, et al. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin Proc*. 2008;83(5):584–594.
- Gu D, Kelly TN, Wu X, et al. Mortality attributable to smoking in China. *N Engl J Med*. 2009;360(2):150–159.
- Mattson ME, Pollack ES, Cullen JW. What are the odds that smoking will kill you? *Am J Public Health*. 1987;77(4):425–431.
- Peto R, Darby S, Deo H, et al. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *BMJ*. 2000;321(7257):323–329.
- Matakidou A, Eisen T, Houlston RS. Systematic review of the relationship between family history and lung cancer risk. *Br J Cancer*. 2005;93(7):825–833.
- Herbst RS, Heymach JV, Lippman SM. Lung cancer. *N Engl J Med*. 2008;359(13):1367–1380.
- Ritchie JC, Idle JR. Population studies of polymorphism in drug oxidation and its relevance to carcinogenesis. *IARC Sci Publ*. 1982;(39):381–394.
- Vineis P, Manuguerra M, Kavvoura FK, et al. A field synopsis on low-penetrance variants in DNA repair genes and cancer susceptibility. *J Natl Cancer Inst*. 2009;101(1):24–36.
- Yu W, Gwinn M, Clyne M, et al. A navigator for human genome epidemiology. *Nat Genet*. 2008;40(2):124–125.
- Dong LM, Potter JD, White E, et al. Genetic susceptibility to cancer: the role of polymorphisms in candidate genes. *JAMA*. 2008;299(20):2423–2436.
- Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor sub-unit genes on 15q25.1. *Nature*. 2008;452(7187):633–637.
- Spinola M, Leoni VP, Galvan A, et al. Genome-wide single nucleotide polymorphism analysis of lung cancer risk detects the *KLF6* gene. *Cancer Lett*. 2007;251(2):311–316.
- Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet*. 2008;40(5):616–622.
- Galvan A, Falvella FS, Spinola M, et al. A polygenic model with common variants may predict lung adenocarcinoma risk in humans. *Int J Cancer*. 2008;123(10):2327–2330.
- Liu P, Vikis HG, Wang D, Lu Y, et al. Familial aggregation of common sequence variants on 15q24-25.1 in lung cancer. *J Natl Cancer Inst*. 2008;100(18):1326–1330.
- Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet*. 2008;40(12):1407–1409.
- McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet*. 2008;40(12):1404–1406.
- Ioannidis JP. Common genetic variants for breast cancer: 32 largely refuted candidates and larger prospects. *J Natl Cancer Inst*. 2006;98(19):1350–1353.
- Chang CH, Hsiao CF, Chang GC, et al. Interactive effect of cigarette smoking with human 8-oxoguanine DNA *N*-glycosylase 1 (*hOGG1*) polymorphisms on the risk of lung cancer: a case-control study in Taiwan. *Am J Epidemiol*. 2009;170(6):695–702.
- Hung RJ, Christiani DC, Risch A, et al. International Lung Cancer Consortium: pooled analysis of sequence variants in DNA repair and cell cycle pathways. *Cancer Epidemiol Biomarkers Prev*. 2008;17(11):3081–3089.
- Li H, Hao X, Zhang W, et al. The *hOGG1 Ser326Cys* polymorphism and lung cancer risk: a meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2008;17(7):1739–1745.
- Botto LD, Khoury MJ. Commentary: facing the challenge of gene-environment interaction: the two-by-four table and beyond. *Am J Epidemiol*. 2001;153(10):1016–1020.
- Wacholder S, Chatterjee N, Caporaso N. Intermediacy and gene-environment interaction: the example of *CHRNA5-A3* region, smoking, nicotine dependence, and lung cancer. *J Natl Cancer Inst*. 2008;100(21):1488–1491.
- Merikangas KR, Risch N. Genomic priorities and public health. *Science*. 2003;302(5645):599–601.

29. Carlsten C, Burke W. Potential for genetics to promote public health: genetics research on smoking suggests caution about expectations. *JAMA*. 2006;296(20):2480–2482.
30. Khoury MJ, Davis R, Gwinn M, et al. Do we need genomic research for the prevention of common diseases with environmental causes? *Am J Epidemiol*. 2005;161(9):799–805.
31. Hecht SS. Cigarette smoking and lung cancer: chemical mechanisms and approaches to prevention. *Lancet Oncol*. 2002;3(8):461–469.
32. Field JK. Lung cancer risk models come of age. *Cancer Prev Res (Phila Pa)*. 2008;1(4):226–228.
33. Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies—challenges and opportunities. *Am J Epidemiol*. 2009;169(2):227–230.
34. Taioli E. Gene-environment interaction in tobacco-related cancers. *Carcinogenesis*. 2008;29(8):1467–1474.
35. Shields PG. Molecular epidemiology of smoking and lung cancer. *Oncogene*. 2002;21(45):6870–6876.
36. Zintzaras E, Lau J. Synthesis of genetic association studies for pertinent gene–disease associations requires appropriate methodological and statistical approaches. *J Clin Epidemiol*. 2008;61(7):634–645.
37. Landi MT, Consonni D, Rotunno M, et al. Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. *BMC Public Health*. 2008;8:203.
38. Khoury MJ, Little J. Human Genome Epidemiologic reviews: the beginning of something HuGE. *Am J Epidemiol*. 2000;151(1):2–3.
39. Wu T, Hu Y, Chen C, et al. Passive smoking, metabolic gene polymorphisms, and infant birth weight in a prospective cohort study of Chinese women. *Am J Epidemiol*. 2007;166(3):313–322.
40. Susser M. Does risk factor epidemiology put epidemiology at risk? Peering into the future. *J Epidemiol Community Health*. 1998;52(10):608–611.
41. Davey Smith G, Ebrahim S. Epidemiology—is it time to call it a day? *Int J Epidemiol*. 2001;30(1):1–11.
42. Ray R, Loughhead J, Wang Z, et al. Neuroimaging, genetics and the treatment of nicotine addiction. *Behav Brain Res*. 2008;193(2):159–169.
43. Vineis P, Kriebel D. Causal models in epidemiology: past inheritance and genetic future [electronic article]. *Environ Health*. 2006;5:21.
44. Pearce N. Commentary: the rise and rise of corporate epidemiology and the narrowing of epidemiology’s vision. *Int J Epidemiol*. 2007;36(4):713–717.